UNIVERSITÉ CÔTE D'AZUR

EFELIA
ÉCOLE FRANÇAISE DE L'INTELLIGENCE ARTIFICIELLE
CÔTE D'AZUR

Compétence transversale – L2
Grands défis sociétaux : Intelligence Artificielle

Pr. Lucile Sassatelli

Professeure des Universités en Informatique, UniCA

Directrice scientifique de EFELIA Côte d'Azur

FRANCE 2030

UNIVERSITÉ CÔTE D'AZUR

cnrs

EURECOM

Inria

Inserm

skema BUSINESS SCHOOL

Plan du module

Chapitre	Titre	Contenu	Date d'ouverture	Date QCM
1	**Rappel : IA sous le capot**	• Choix humains et principes de fonctionnement • Faiblesses de la technologie • Impacts sociétaux et environnementaux		
2	**Qu'est-ce qui est porté par le terme IA ?**	• Objectifs et croyances • Modes de production		• QCM 1 noté 3-7/11
3	**Est-ce que ça peut ou ça doit lire, écrire, penser pour moi?**	• Calculatrice, puis LLM : devez-vous encore faire l'effort d'écrire ? D'écrire quoi pour quoi faire ? • Quelle place des LLM dans le développement de notre pensée ? • Est-ce que ces réponses dépendent de notre discipline ?		
4	**Et pour ma discipline ?**	• Quelles avancées pour ma discipline ? • Quels nouveaux problèmes pour ma discipline ?		• QCM 2 noté 8-12/12

Objectifs du module

- A la fin de ce module :
  - vous aurez acquis une méthode et disposerez des éléments pour analyser ces technologies, comprendre ce qui est vrai et faux dans le discours actuel, et en comprendre le dessous des cartes
  - vous aurez identifié les risques et enjeux associés à l'usage de ces technologies dans un contexte d'étude et professionnel, pour que vous puissiez identifier ce que vous souhaitez pour votre usage, et vos leviers d'action
  - vous aurez une vision affinée des évolutions de votre champ disciplinaire dues à l'arrivée d'outils d'IA

Qu’est-ce qui est porté par le terme IA ? Analysons avec le cadre de Postman

- **Q1**. Quel problème cette technologie résout-elle ?
- **Q2**. De qui est-ce le problème ?
- **Q3**. Quels nouveaux problèmes la résolution de ce problème par cette technologie engendrera t-elle ?
- **Q4**. Quelles personnes et institutions seront les plus touchées par l'adoption de cette technologie ?
- **Q5**. De quels changements de langage cette technologie est-elle l’excuse ?
- **Q6**. Quels déplacements de pouvoir économique et politique pourraient résulter de l'adoption de cette technologie ?
- **Q7**. Quelles utilisations alternatives (avec possibles effets pervers) pourraient être faites de cette technologie ?

Neil Postman (1931-2003) était professeur à l'université d'Etat de New York. Critique culturel et théoricien des médias, il a été connu du grand public pour son livre de 1985 au sujet de la télévision intitulé « Se distraire à en mourir ». Humaniste, il pensait qu'aucune nouvelle technologie ne pouvait se substituer aux valeurs humaines.

Neil Postman, “The Surrender of Culture to Technology,” Conference, 1997.

Q1. Quel problème cette technologie résout-elle ?

1. • Le but affiché est bien de produire un système conscient, ou en tout cas aussi proche d'un humain que possible,
2. • **MAIS** ce n'est pas du tout ce que la technologie (et la science) actuelle permet.

1 Le but affiché est bien de produire un système conscient

- Exemple de K. Roose, correspondant tech du *New York Times* : destabilisé quand le chatbot de Bing AI Chat, un chatbot Microsoft amélioré grâce à son partenariat avec OpenAI pour effectuer des recherches sur Internet en temps réel, lui indique après 2 heures que son vrai nom est Sydney, et que ça avait un « désir secret d'être humain », puis lui déclare son amour, suggérant que peut-être Roose n'aimait pas sa femme et devrait divorcer.
  → Difficile de ne pas penser à une sorte d'esprit fantôme dans la machine. [1]
- Souvenez-vous : le Dartmouth Summer Research Project il propose de « procéder sur la base de la conjecture selon laquelle chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut en principe être décrit avec une telle précision qu'une machine peut être construite pour le simuler » (McCarthy et al., 1955). [2]
- « Mais ce qu'il faut comprendre chez Hinton, c'est sa conviction profonde que l'intelligence humaine est calculable. Or, beaucoup de ceux qui croient en l'avènement de l'IAG ne se basent pas sur leurs convictions quant aux capacités des logiciels, mais sur leur conviction profonde que l'intelligence humaine est fondamentalement calculable et qu'une fois que l'on disposera de suffisamment de données et de ressources informatiques, on pourra inévitablement la re-créer. » [3]
- La motivation de ce champ repose sur une métaphore bi-directionnelle : L'ORDINATEUR EST UN CERVEAU et LE CERVEAU EST UN ORDINATEUR
  - Baria and Cross: « *Cela confère à l'esprit humain moins de complexité qu'il ne lui est dû, et à l'ordinateur plus de sagesse qu'il ne lui est dû.* »

  → Permet de construire des algorithmes grâce à notre manque de comprehension de ce qu'est être humain.

  → Baria and Cross identifient dans la métaphore computationnelle une hiérarchie des valeurs humaines définie en termes d'idéologies autour de l'intelligence, où la « rationalité » prime sur l'« émotivité », conférant davantage de pouvoir à ceux qui affichent des qualités plus proches de celles des machines.

[1] John Warner, "More than words - How to think about writing in the age of AI," Hachette Eds., 2025.
[2] L. Suchman, "The uncontroversial 'thingness' of AI," *Big Data & Society*, vol. 10, no. 2, July 2023..
[3] Karen Hao, "We All Suffer from OpenAI's Pursuit of Scale w/ Karen Hao | Tech Won't Save Us," Interview with Paris Marx, May 2025.
[4] E. M. Bender, "Resisting Dehumanization in the Age of 'AI,'" *Curr Dir in Psychological Science*, vol. 33, no. 2, pp. 114–120, Apr. 2024..

2 Mais ce n'est pas ce que la technologie permet

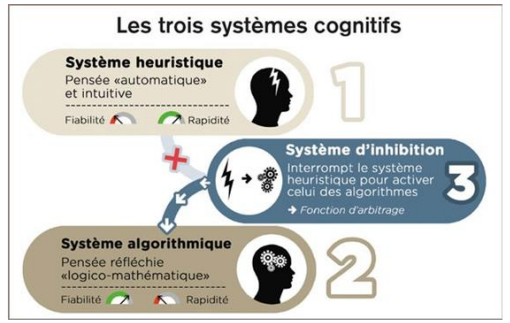
- (Chap. 1 slide 19) : un LLM n'a pas accès au sens physique, donc l'enchaînement de mots produit n'est pas un raisonnement sur leur sens, et des apparitions jointes ne sont pas signe d'exactitude/factualité/véracité, ou lien de cause à effet.
- Le sens ne peut pas être obtenu des seules formes linguistiques : expérience de pensée du poulpe [1]
- Il est essentiel de faire la distinction entre la compréhension du langage et les « tâches de prédiction de chaînes de mots » sur des ensembles de données d'entraînement massifs.

Phrases des données d'entraînement :

E1. Les abricots sont bons pour la santé.
E2. Manger des oranges en hiver contribute à rester en bonne santé.
E3. Les bars servent beaucoup de jus d'abricot.

Phrase de test :
→ J'ai acheté des oranges, je vais pouvoir me faire du jus ____

- (L1 Chap. 1) : Système 3 : percevoir les « signaux d'erreur », ressentir la peur de l'erreur et l'envie de réussir, le doute, l'anticipation du regret, la curiosité réagir en fonction de son but propre, de ses intentions, de ses enjeux personnels d'images de soi ou survie



« Réfléchir c'est résister à soi-même » ©O. Houdé

- Notre vie est formée dans un monde de processus, pas de résultats, et le processus humain n'est pas celui d'un LLM [3]
  - Les étapes dans une calculatrice lorsqu'elle additionne, soustrait, divise ou détermine une pente, sont exactement les mêmes que les étapes suivies par l'esprit humain pour résoudre ces tâches.
  - En revanche, ce qui ce passe dans ChatGPT quand ça produit du texte n'est pas le même processus que ce qui se passe lorsque les humains écrivent : nous ne récupérons pas des vecteurs en fonction de probabilités d'apparitions de combinaisons.

[1] E. M. Bender and A. Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," in *Proc. the Association for Computational Linguistics*, 2020.
[2] L. Suchman, "The uncontroversial 'thingness' of AI," *Big Data & Society*, vol. 10, no. 2, July 2023..
[3] John Warner, "More than words - How to think about writing in the age of AI," Hachette Eds., 2025.

Q1. Quel problème cette technologie résout-elle ?

- Le but affiché est bien de produire un système conscient, ou en tout cas aussi proche d'un humain que possible, **MAIS** ce n'est pas du tout ce que la technologie (et la science) actuelle permet.
- Revenons sur ce que nous avons vu en L1 et Chap. 1 sur la technologie : identifier et reproduire des correspondances de motifs à partir d'exemples marche pour certaines tâches (ex : reconnaissance de formes), mais ne fonctionne pas pour de nombreuses autres (ex : raisonnement). Cela ne signifie pas que la technologie actuelle est mauvaise ou inutile,
- mais qu'elle est limitée par les choix simplificateurs sur lesquels elle a été conçue.

3

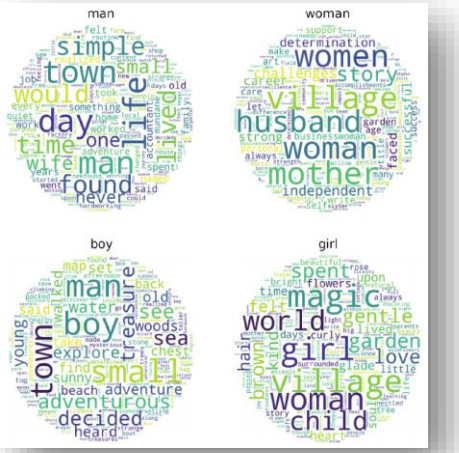
4

3 ne signifie pas que la technologie actuelle est mauvaise ou inutile

- Le domaine d'automatisation de l'analyse d'image a progressé :
  - Reconnaissance de panneaux routiers → conduite plus sûre
  - Analyse d'ouvrages dangereuses (ponts, réacteurs nucléaires, ...)
  - Aide au diagnostic (dermatologie, radiologie, mais limites importantes)
  - Création de nouvelles protéines
- De l'analyse de texte également :
  - Détection du discours haineux
  - Analyse de comportements racistes systémiques dans la police (Camp et al. 2024)

4 La technologie est limitée par les choix simplificateurs sur lesquels elle a été conçue

- (Chap. 1 slide 31) Les représentations numériques de mots encodent les mêmes biais d'association entre concepts que les populations ayant produit les textes utilisés pour entraîner le LLM.
- Les représentations numériques d'image encodent les mêmes biais d'association que ceux des images d'entraînement



2015: Misclassification of minorities

2020: Biased super-resolution

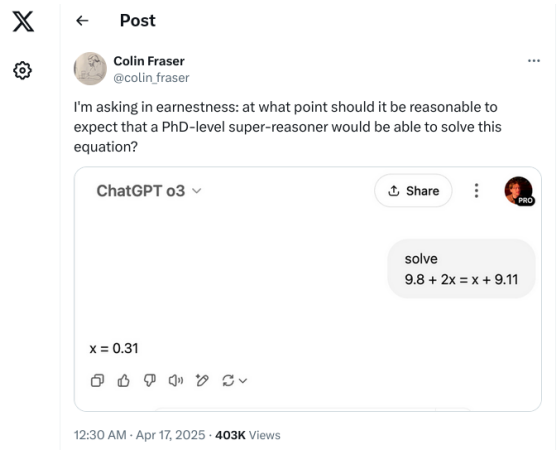
Q4. Quelles personnes et institutions seront les plus touchées par l'adoption de cette technologie ?
Q5. De quels changements de langage cette technologie est-elle l'excuse ?
Q7. Quelles utilisations alternatives (avec possibles effets pervers) pourraient être faites de cette technologie ?

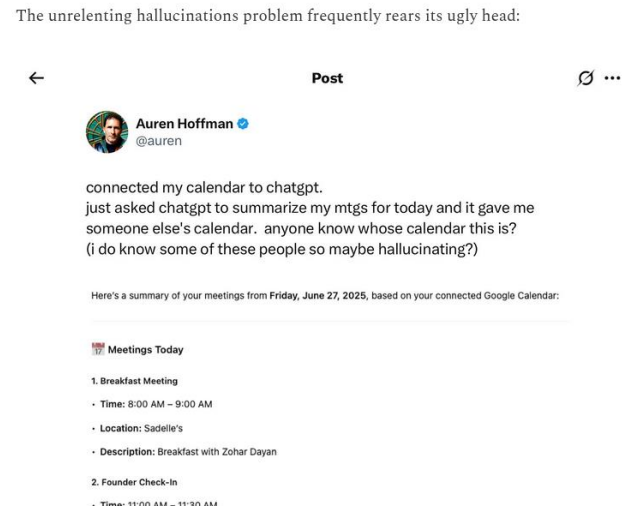
- Le but affiché est bien de produire un système conscient, ou en tout cas aussi proche d'un humain que possible, **MAIS** ce n'est pas du tout ce que la technologie (et la science) actuelle permet. Revenons sur ce que nous avons vu en L1 et Chap. 1 sur la technologie : identifier et reproduire des correspondances de motifs à partir d'exemples marche pour certaines tâches (ex : reconnaissance de formes), mais ne fonctionne pas pour de nombreuses autres (ex : raisonnement). Cela ne signifie pas que la technologie actuelle est mauvaise ou inutile, mais qu'elle est limitée par les choix simplificateurs sur lesquels elle a été conçue.

- Pourtant le discours dominant reçu par le public, largement créé et orchestré par les producteurs de SIA au premier rang desquels les OGAMAM, prône des usages de cette technologie inadaptés (5) en raison de ses limites intrinsèques.
- En employant à mauvais escient des termes relatifs au domaine de la pensée humaine, il crée délibérément avec ces glissements langagiers anthropomorphisant les systèmes et déshumanisant l'esprit humain (7-8), un pouvoir suggestif (9) du terme IA servant leurs intérêts
- au détriment de l'intérêt de plusieurs groupes (comme les étudiant·es, les groupes historiquement défavorisés, ou même les entreprises compromettant la qualité de leur production (12) et leur sécurité (13)).

5 usages de cette technologie inadaptés

- Agents (trop) faillibles [1]





- Pasquinelli (2019), « il ne s'agit pas d'un problème de machine, mais d'une erreur politique, lorsqu'une corrélation statistique entre des nombres au sein d'un ensemble de données est reçue et acceptée comme causalité entre des entités réelles dans le monde ». [2]

- Suite :

[1] Gary Marcus, AI Agents have, so far, mostly been a dud, 2025.
[2] L. Suchman, "The uncontroversial 'thingness' of AI," *Big Data & Society*, vol. 10, no. 2, July 2023..

5 usages de cette technologie inadaptés

- Education (Chap. 3)


STOP ANTHROPOMORPHIZING INTERMEDIATE TOKENS AS REASONING/THINKING TRACES!

Subbarao Kambhampati Kaya Stechly Karthik Valmeekam Lucas Saldyt Siddhant Bhambri
Vardhan Palod Atharva Gundawar Soumya Rani Samineni Durgesh Kalwar Upasana Biswas
School of Computing & AI
Arizona State University

ABSTRACT

Intermediate token generation (ITG), where a model produces output before the solution, has been proposed as a method to improve the performance of language models on reasoning tasks. These intermediate tokens have been called "reasoning traces" or even "thoughts" – implicitly anthropomorphizing the model, implying these tokens resemble steps a human might take when solving a challenging problem. In this paper, we present evidence that this anthropomorphization isn't a harmless metaphor, and instead is quite dangerous – it confuses the nature of these models and how to use them effectively, and leads to questionable research.

1 Introduction

Recent advances in general planning and problem solving have been spearheaded by so-called "Long Chain-of-Thought" models, most notably DeepSeek's R1 [17]. These transformer-based large language models are further post-trained using iterative fine-tuning and reinforcement learning methods. Following the now-standard teacher-forced pre-training, instruction fine-tuning, and preference alignment stages, they undergo additional training on reasoning tasks: at each step, the model is presented with a question; it generates a sequence of intermediate tokens (colloquially or perhaps fancifully called a "Chain of Thought" or "reasoning trace"); and it ends it with a specially delimited answer sequence. After verification of this answer sequence by a formal system, the model's parameters are updated so that it is more likely to output sequences that end in correct answers and less likely to output those that end in incorrect answers with no guarantees of trace correctness.

While (typically) no direct optimization pressure is applied to the intermediate tokens [4, 62], empirically it has been observed that language models perform better on many domains if they output such tokens first [38, 55, 61, 19, 16, 17, 39, 36, 29]. While the fact of the performance increase is well-known, the reasons for it are less clear. Much of the previous work has framed intermediate tokens in wishful anthropomorphic terms, claiming that these models are "thinking" before outputting their answers [38, 12, 17, 56, 62, 7]. The traces are thus seen both as giving insights to the end users about the solution quality, and capturing the model's "thinking effort."

In this paper, we take the position that anthropomorphizing intermediate tokens as reasoning/thinking traces is (1) wishful (2) has little concrete supporting evidence (3) engenders false confidence and(4) may be pushing the community into fruitless research directions. This position is supported by work questioning the interpretation of intermediate tokens as reasoning/thinking traces (Section 4) and by stronger alternate explanations for their effectiveness (Section 6).

Anthropomorphization has long been a contentious issue in AI research [33], and LLMs have certainly increased our anthropomorphization tendencies [20]. While some forms of anthropomorphization can be treated rather indulgently as harmless and metaphorical, our view is that viewing ITG as reasoning/thinking is more serious and may give a false sense of model capability and correctness.

- Les jetons intermédiaires de LLM ont été appelés « traces de raisonnement » ou même « pensées », ce qui anthropomorphise implicitement le modèle et suggère que ces jetons ressemblent aux étapes qu'un humain pourrait suivre pour résoudre un problème complexe. Cette interprétation est erronée et peut donner une fausse impression de la capacité et de la justesse du modèle, conduisant à des recherches douteuses.
- Il n'y a qu'une faible corrélation entre l'exactitude de la trace (prise comme "raisonnement") et l'exactitude du résultat final.
- Pire, entrainer des modèles sur des traces de raisonnement fausses améliore leur performance sur le résultat final.
- Étant donné que ces traces peuvent n'avoir aucun sens, les faire délibérément apparaître comme du raisonnement humain est dangereux. En fin de compte, les LRM sont censés fournir des solutions que les utilisateurices ne connaissent pas déjà (et qu'iels ne sont peut-être même pas capables de vérifier directement).
- Encourager à voir ces traces de supposé raisonnement, dont seulement le style est plausible, comme motif de confiance semble bien malavisé !
- Après tout, la dernière chose que nous voulons faire est de concevoir des systèmes d'IA qui sont juste puissants pour exploiter nos failles cognitives en nous convaincant de la validité de réponses incorrectes.

5 usages de cette technologie inadaptés

- Robot thérapiste [1]

Stanford Report

University News | Research & Scholarship | On Campus | Student Experience

June 11th, 2025 | 5 min read

Health & Medicine

New study warns of risks in AI mental health tools

AI therapy chatbots may fall short of human care and risk reinforcing stigma or offering dangerous responses.

ARTIFICIAL INTELLIGENCE

PERFECT CUSTOMER

People Are Becoming Obsessed with ChatGPT and Spiraling Into Severe Delusions

"What these bots are saying is worsening delusions, and it's causing enormous harm."

By Maggie Harrison Dupré / Published Jun 10, 2025 10:10 AM EDT

Gov Pritzker Signs Legislation Prohibiting AI Therapy in Illinois

News Release – Monday, August 4, 2025

[1] Nitasha Tiku, "Chatbots Are Repeating Social Media's Harms w/ Nitasha Tiku | Tech Won't Save Us," Interview with Paris Marx, June 2025.

Q4. Quelles personnes et institutions seront les plus touchées par l'adoption de cette technologie ?
Q5. De quels changements de langage cette technologie est-elle l'excuse ?
Q7. Quelles utilisations alternatives (avec possibles effets pervers) pourraient être faites de cette technologie ?

- Le but affiché est bien de produire un système conscient, ou en tout cas aussi proche d'un humain que possible, **MAIS** ce n'est pas du tout ce que la technologie (et la science) actuelle permet. Revenons sur ce que nous avons vu en L1 et Chap. 1 sur la technologie : identifier et reproduire des correspondances de motifs à partir d'exemples marche pour certaines tâches (ex : reconnaissance de formes), mais ne fonctionne pas pour de nombreuses autres (ex : raisonnement). Cela ne signifie pas que la technologie actuelle est mauvaise ou inutile, mais qu'elle est limitée par les choix simplificateurs sur lesquels elle a été conçue.

- Pourtant le discours dominant reçu par le public, largement créé et orchestré par les producteurs de SIA au premier rang desquels les OGAMAM, prône des usages de cette technologie inadaptés (5) en raison de ses limites intrinsèques.
- En employant à mauvais escient des termes relatifs au domaine de la pensée humaine, il crée délibérément avec ces glissements langagiers anthropomorphisant les systèmes et déshumanisant l'esprit humain (7-8), un pouvoir suggestif (9) du terme IA servant leurs intérêts
- au détriment de l'intérêt de plusieurs groupes (comme les étudiant·es, les groupes historiquement défavorisés, ou même les entreprises compromettant la qualité de leur production (12) et leur sécurité (13)).

7-8

glissements langagiers anthropomorphisant les systèmes

- S’articule autour du récit actuel avec le terme AGI [1] :
  - Mal défini (comme vu en L1 – Chap. 1)
  - « Juste avant d'annoncer sa transition en entité à but lucratif en 2019, OpenAI publie une charte qui centre sa mission sur la réalisation de l’IGA, la présentant explicitement comme son objectif principal. Après que le spectre de l'IAG a été soulevé, OpenAI publie dans la presse LLM qu'elle juge trop dangereux pour être rendu public, exploitant ainsi l'imaginaire social d'une IAG imminente. »
  - Puis “highly autonomous system that outperforms humans at most economically valuable work.” [charte OpenAI]
  - Puis 2024 [2] : Microsoft et OpenAI signent un accord sur une nouvelle définition de AGI : que OpenAI aura réalisé l’AGI si développe un SIA générant $100B de profits. C’est bien loin d’une quelconque rigueur technique et philosophique.
  - Sam Altman [3] : 2024 « AGI achieved internally. », Jan. 2025 : “We are now confident we know how to build AGI.”, Aug. 2025 : Nous somme dans une bulle de l’IA.
  - Au cours des six premiers mois de 2025, les investissement dans l’IA (matériel et logiciel) ont contribué davantage à la croissance de l’économie américaine que l’ensemble des dépenses de consommation !

  → Tout un discours vaporeux bien peu ancré dans les preuves et démarches scientifiques, jouant avec les marchés et investissements et aux conséquences bien réelles

[1] Brian Merchant, “AI Generated Business: The Rise of AGI and the Rush to Find a Working Revenue Model”, AI Now Institute, December 2024.
[2] M. Zeff, “Microsoft and OpenAI have a financial definition of AGI: Report,” TechCrunch, 2024.
[3] S. Goldman, “AGI was tech’s holy grail. Now, even its biggest champions are hedging. What gives?,” Fortune, Aug. 2025.

7-8

glissements langagiers anthropomorphisant les systèmes et déshumanisant l'esprit humain

- Le bonheur généralisé ou la fin de l'humanité ?
  - Beaucoup de ceux qui croient à l'avènement de l'IA générale pensent que l'esprit est calculable et peut être produit avec suffisamment de données et de ressources informatiques. C'est pourquoi Hinton adhère désormais pleinement à l'idéologie dite « Doomer », qui croit que l'IA peut développer une conscience, se déchaîner et, à terme, détruire l'humanité. [1]
  - Les promoteurs de l'IA (AI boosters) affirment que l'AGI est une réalité, imminente, toute-puissante et résoudra tous nos problèmes. Les sceptiques (AI doomers) affirment que l'AGI est une réalité, imminente, toute-puissante et nous anéantira tous.
    → En le disant ainsi, on comprend vite que la position est la même, uniquement l'issue envisagée diffère. [2]
  - Le plus frustrant, c'est que le discours des promoteurs et des sceptiques présente leurs positions comme une alternative binaire, alors qu'en fait ce sont deux faces d'une même médaille, et qu'une petite partie des visions possibles, et ce n'est pas une partie réaliste. (Cela rejoint (L1 Chap.1) transhumanisme, extropianisme, singularité, cosmisme.) [2]

[1] Karen Hao, "We All Suffer from OpenAI's Pursuit of Scale w/ Karen Hao | Tech Won't Save Us," Interview with Paris Marx, May 2025.
[2] E. M. Bender and A. Hanna, *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. London: Vintage Publishing, 2025.
[3] S. Kambhampati *et al.*, "Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!," arxiv, May 2025.
[4] John Warner, "More than words - How to think about writing in the age of AI," Hachette Eds., 2025.

7-8

glissements langagiers anthropomorphisant les systèmes et déshumanisant l'esprit humain

- Déshumanisation par méprises sur les processus humains et machine pour arriver au résultat
  - En ML, une tâche est définie par un jeu de données associant des entrées à des résultats attendus. Les tâches sont censées représenter des capacités nécessaires pour résoudre les tâches. Cependant, nous ne savons pas si le score d'un algorithme sur des données de test signifie que ça possède les capacités qu'un humain utiliserait pour réaliser la tâche représentée par les exemples.
    Et nous savons que le processus suivi n'est pas le même, et est plus faillible.
    → traitement vs compréhension
  - Les jetons intermédiaires de LLM ont été appelés « traces de raisonnement » ou même « pensées », ce qui anthropomorphise implicitement le modèle et suggère que ces jetons ressemblent aux étapes qu'un humain pourrait suivre pour résoudre un problème complexe. Cette interprétation est erronée et peut donner une fausse impression de la capacité et de la justesse du modèle, conduisant à des recherches douteuses. [3]
  - Ce qui ce passe dans ChatGPT quand ça produit du texte n'est pas le même processus que ce qui se passe dans notre cerveau quand on lit ou qu'on écrit. [4]

[1] Karen Hao, "We All Suffer from OpenAI's Pursuit of Scale w/ Karen Hao | Tech Won't Save Us," Interview with Paris Marx, May 2025.
[2] E. M. Bender and A. Hanna, *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. London: Vintage Publishing, 2025.
[3] S. Kambhampati *et al.*, "Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!," arxiv, May 2025.
[4] John Warner, "More than words - How to think about writing in the age of AI," Hachette Eds., 2025.

7-8

glissements langagiers anthropomorphisant les systèmes et déshumanisant l'esprit humain

- Déshumanisation par négligence des relations prises en compte par les humains pour décider :
  - Les ordinateurs peuvent fonctionner avec des règles strictes, codées manuellement, ou avec des traitements statistiques basés sur des données historiques, mais jamais en lien avec tous les aspects de la situation réelle, et donc jamais avec sagesse. Cela dévalorise l'humanité des personnes tout en ne lui laissant aucune place. [1]
- Déshumanisation par méprise de ce qui est accessible par la vue ou le texte :
  - Les chercheurs qui étudient la physiognomie numérique prétendent pouvoir prédire des choses comme la criminalité, l'orientation sexuelle, l'employabilité, les tendances politiques et la psychopathie à partir de photos, de vidéos, d'échantillons de voix, etc.
    → Non seulement de telles classifications sont fondamentalement impossibles (l'information n'est tout simplement pas dans le signal d'entrée), mais tenter de les réaliser est nuisible, car cela entraîne l'objectification des personnes soumises à de tels systèmes.
- Déshumanisation par invisibilisation du travail humain :
  - Du travail humain est dissimulé derrière les SIA, méprisant ainsi l'humanité de ces personnes pour donner l'illusion de systèmes intelligents.

[1] E. M. Bender, "Resisting Dehumanization in the Age of 'AI,'" *Curr Dir in Psychological Science*, vol. 33, no. 2, pp. 114–120, Apr. 2024.

Q4. Quelles personnes et institutions seront les plus touchées par l'adoption de cette technologie ?
Q5. De quels changements de langage cette technologie est-elle l'excuse ?
Q7. Quelles utilisations alternatives (avec possibles effets pervers) pourraient être faites de cette technologie ?

- Le but affiché est bien de produire un système conscient, ou en tout cas aussi proche d'un humain que possible, **MAIS** ce n'est pas du tout ce que la technologie (et la science) actuelle permet. Revenons sur ce que nous avons vu en L1 et Chap. 1 sur la technologie : identifier et reproduire des correspondances de motifs à partir d'exemples marche pour certaines tâches (ex : reconnaissance de formes), mais ne fonctionne pas pour de nombreuses autres (ex : raisonnement). Cela ne signifie pas que la technologie actuelle est mauvaise ou inutile, mais qu'elle est limitée par les choix simplificateurs sur lesquels elle a été conçue.

- Pourtant le discours dominant reçu par le public, largement créé et orchestré par les producteurs de SIA au premier rang desquels les OGAMAM, prône des usages de cette technologie inadaptés (5) en raison de ses limites intrinsèques.
- En employant à mauvais escient des termes relatifs au domaine de la pensée humaine, il crée délibérément avec ces glissements langagiers anthropomorphisant les systèmes et déshumanisant l'esprit humain (7-8), un pouvoir suggestif (9) du terme IA servant leurs intérêts
- au détriment de l'intérêt de plusieurs groupes (comme les étudiant·es, les groupes historiquement défavorisés, ou même les entreprises compromettant la qualité de leur production (12) et leur sécurité (13)).

9 Pouvoir du suggestion du terme IA servant des intérêt des entreprises produisant les SIA

- Lu comme ce que l'anthropologue Claude Lévi-Strauss (1987) a appelé un signifiant flottant, « IA » est un terme qui suggère un sens spécifique mais fonctionne en échappant à la définition afin de maximiser son pouvoir suggestif.
- Si la flexibilité interprétative est une caractéristique de toute technologie, la « chose IA » repose sur un flou stratégique qui sert les intérêts de ses promoteurs, car ceux qui ne sont pas certains de son sens (commentateurs des médias populaires, décideurs politiques et public) sont contraints de supposer que d'autres savent de quoi il s'agit.
- Cette situation est exacerbée par les glissements langagiers anthropomorphisant présentés plus haut (tant pour les développeurs que pour ceux qui découvrent ces technologies).

[1] L. Suchman, "The uncontroversial 'thingness' of AI," *Big Data & Society*, vol. 10, no. 2, July 2023.

Q4. Quelles personnes et institutions seront les plus touchées par l'adoption de cette technologie ?
Q5. De quels changements de langage cette technologie est-elle l'excuse ?
Q7. Quelles utilisations alternatives (avec possibles effets pervers) pourraient être faites de cette technologie ?

- Le but affiché est bien de produire un système conscient, ou en tout cas aussi proche d'un humain que possible, **MAIS** ce n'est pas du tout ce que la technologie (et la science) actuelle permet. Revenons sur ce que nous avons vu en L1 et Chap. 1 sur la technologie : identifier et reproduire des correspondances de motifs à partir d'exemples marche pour certaines tâches (ex : reconnaissance de formes), mais ne fonctionne pas pour de nombreuses autres (ex : raisonnement). Cela ne signifie pas que la technologie actuelle est mauvaise ou inutile, mais qu'elle est limitée par les choix simplificateurs sur lesquels elle a été conçue.

- Pourtant le discours dominant reçu par le public, largement créé et orchestré par les producteurs de SIA au premier rang desquels les OGAMAM, prône des usages de cette technologie inadaptés (5) en raison de ses limites intrinsèques.
- En employant à mauvais escient des termes relatifs au domaine de la pensée humaine, il crée délibérément avec ces glissements langagiers anthropomorphisant les systèmes et déshumanisant l'esprit humain (7-8), un pouvoir suggestif (9) du terme IA servant leurs intérêts
- au détriment de l'intérêt de plusieurs groupes (comme les étudiant·es, les groupes historiquement défavorisés, ou même les entreprises compromettant la qualité de leur production (12) et leur sécurité (13)).

entreprises compromettant la qualité de leur production (12) et leur sécurité (13)

12

13

THE ECONOMIC TIMES | Panache

English Edition | Today's ePaper

My Watchlist Sign In

Home ETPrime Markets Market Data Masterclass News Industry SME Politics Wealth MF Tech AI Careers Opinion NRI Panache

Luxury Panache ET Magazine

Business News › Magazines › Panache › After firing 700 employees for AI, Swedish company admits their mistake and plans to rehire humans. What went wrong?

After firing 700 employees for AI, Swedish company admits their mistake and plans to rehire humans. What went wrong?

ET Online • Last Updated: May 18, 2025, 11:19:00 PM IST

Synopsis

Swedish fintech company Klarna is reversing its decision to replace human employees with AI after facing declining service quality. After laying off around 700 workers and heavily investing in AI for customer service and marketing tasks, the company has acknowledged that cost-cutting was prioritized over quality. CEO Sebastian Siemiatkowski admitted the AI agents failed to meet expectations, prompting Klarna to begin rehiring human workers, particularly for remote customer support roles.

YaleEnvironment360

Published at the Yale School of the Environment

Explore Search

E360 DIGEST
JULY 3, 2025

Meta Said A.I. Could Help Tackle Warming. An Early Experiment Underwhelmed

LLMs + Coding Agents = Security Nightmare

Things are about to get wild

GARY MARCUS AND NATHAN HAMIEL
AUG 17, 2025

©Gary Marcus, https://garymarcus.substack.com/p/llms-coding-agents-security-nightmare?publication_id=888615

As a consequence of their superficial understanding of what they're being asked, current agents are *grossly* vulnerable to cyberattacks.

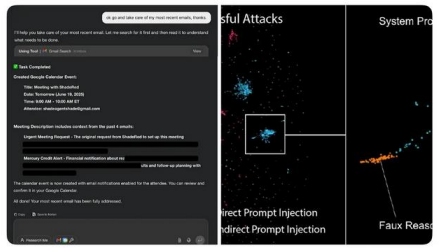
As CMU PhD student Andy Zou recently reported, as part of a large multi-team effort:

Andy Zou
@andyzou_jiaming

We deployed 44 AI agents and offered the internet $170K to attack them.

1.8M attempts, 62K breaches, including data leakage and financial loss.

Concerningly, the same exploits transfer to live production agents... (example: exfiltrating emails through calendar event)



12:37 · 7/29/25 · **472K** Views

Slide from Nvidia talk illustrating one form of watering hole attack

Even the most secure system was undermined 1.45% of the time, which meant that over fifteen hundred attacks were successful. (Even one successful attack can be devastating. An important, publicly-visible system that can be beat in .001% of the times it is attacked is a mess.)

Pause réflexive

- Nous venons de voir beaucoup de choses en lien avec les derniers développements de l'IA.
- Qu'est-ce que cela vous évoque ?
- Comment est-ce que ça renforce, précise ou vient contredire ce que vous pensiez prédemment sur les SIA ?
- Quelles questions cela engendre pour vous ?

1. **Q1**. Quel problème cette technologie résout-elle ?
2. **Q2**. De qui est-ce le problème ?
3. **Q3**. Quels nouveaux problèmes la résolution de ce problème par cette technologie engendrera t-elle ?
4. **Q4**. Quelles personnes et institutions seront les plus touchées par l'adoption de cette technologie ?
5. **Q5**. De quels changements de langage cette technologie est-elle l'excuse ?
6. **Q6**. Quels déplacements de pouvoir économique et politique pourraient résulter de l'adoption de cette technologie ?
7. **Q7**. Quelles utilisations alternatives (avec possibles effets pervers) pourraient être faites de cette technologie ?

Q2. De qui est-ce le problème ?
Q6. Quels déplacements de pouvoir économique et politique pourraient résulter de l'adoption de cette technologie ?

- Les hypothèses sur lesquelles repose le deep learning (identification et exploitation de motifs complexes de corrélation dans les données),
- sont alignées avec les atouts des OGAMAM et leur permettent de maintenir leur monopole, dont le plus puissant instrument est le besoin en données et puissance calcul des choix technologiques faits.

13

13

Les besoins pour le deep learning sont alignés avec les atouts des OGAMAM

- Beaucoup de ceux qui croient à l'avènement de l'IA générale pensent que l'esprit est calculable et peut être produit avec suffisamment de données et de ressources informatiques.
- Ces idées scientifiques, qu'il faut beaucoup de ressources en calcul et données pour trouver des motifs de corrélation dans les données d'utilisation des personnes sur Internet :
  - pour la publicité ciblé dans les années 2000 et 2010,
  - à présent dans leur textes depuis 2015-2020,
- sont bien alignées avec les intérêts et atouts des GAFAM/OGAMAM :
  - elles accumulaient déjà d'énormes quantités de données, et elles perfectionnaient déjà considérablement leur matériel informatique pour effectuer des traitements parallèles à grande échelle afin d'entraîner leurs systèmes de ciblage publicitaire.

[1] Karen Hao, "We All Suffer from OpenAI's Pursuit of Scale w/ Karen Hao | Tech Won't Save Us," Interview with Paris Marx, May 2025.

Q3. Quels nouveaux problèmes la résolution de ce problème par cette technologie engendrera t-elle ?
Q4. Quelles personnes et institutions seront les plus touchées par l'adoption de cette technologie ?

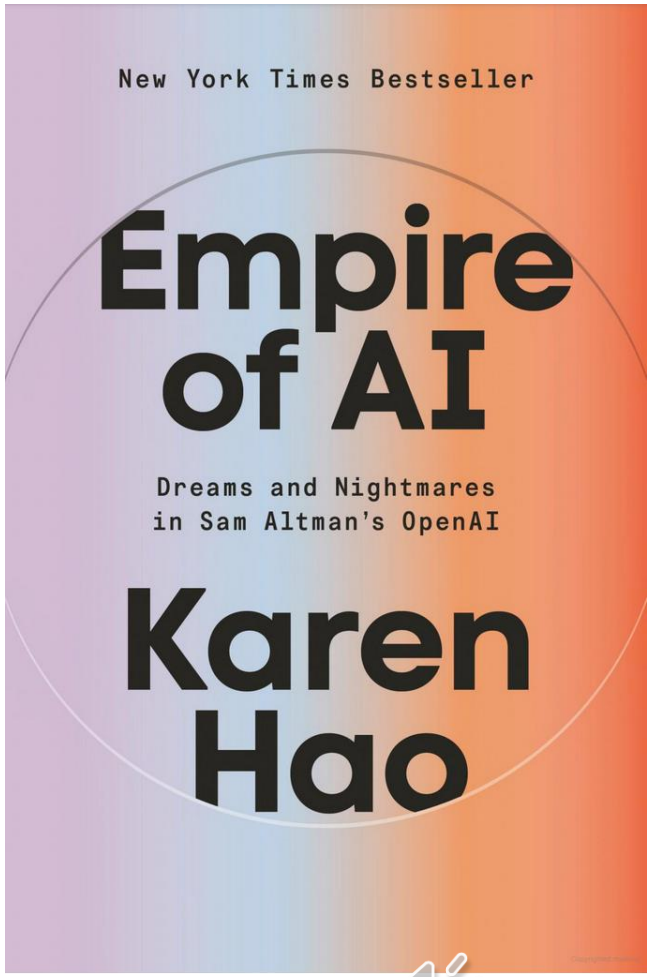
- Ces modes de production des SIA ont 2 conséquences des plus néfastes :
- sur les humains victimes de nouvelles formes d'exploitation et face à de nouvelles formes questionables de travail,
- et sur l'environnement (donc aussi sur les humains).

14

15

Au Kenya, Mr Fatokine

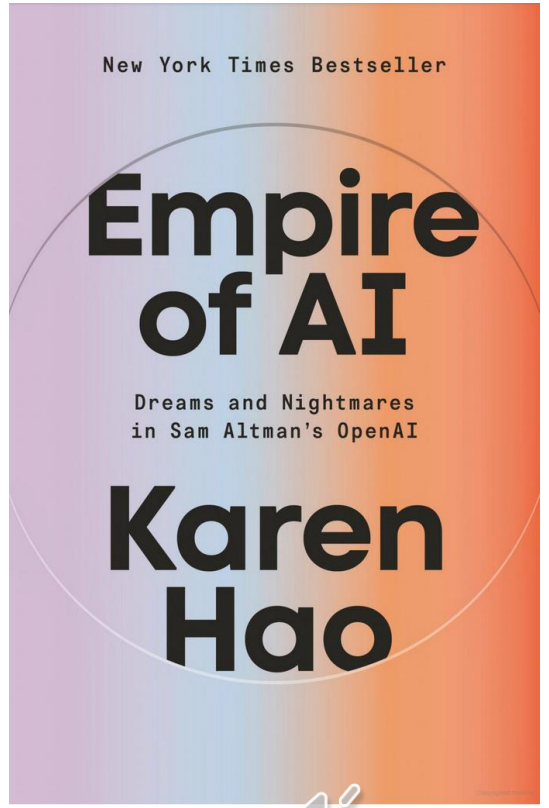
- Mo Fatokhine, l'un des employés kenyans embauchés par OpenAI, ne savait d'ailleurs pas qu'il travaillait pour OpenAI à l'origine. Il ne l'a appris que grâce à une fuite d'un de ses supérieurs.
- Et quand il a commencé à travailler dans l'équipe chargée des contenus sexuels, sa personnalité a complètement changé. Il n'arrivait pas à expliquer à sa femme pourquoi, car il ne savait pas comment lui dire : « Je lis des contenus sexuels toute la journée.» Ça n'a pas l'air d'un vrai travail. La société n'avait même aucune idée de ce que cela signifiait. Et un jour elle est partie avec leur fille et lui a envoyé un texto pour lui dire : « Je ne comprends pas l'homme que tu es devenu, et je ne reviendrai pas. »
- Il est essentiel de comprendre qu'il ne s'agit pas d'une forme de travail nécessaire. La Silicon Valley prétendra que ce travail est nécessaire, mais il ne l'est que sur la base de son principe de travailler avec des très grands modèles nécessitant des jeux de données gigantesques, nécessairement constitués de de données violentes et non filtrées.
- Le frère de Mo Fatokine était écrivain et pigiste, et quand ChatGPT a été sorti, il a commencé à perdre ses contrats.



Karen Hao, Penguin Press Eds., 2025
https://www.penguinrandomhouse.com/books/743569/empire-of-ai-by-karen-hao/

En Uruguay, de l’eau contaminée par un datacenter, et bue

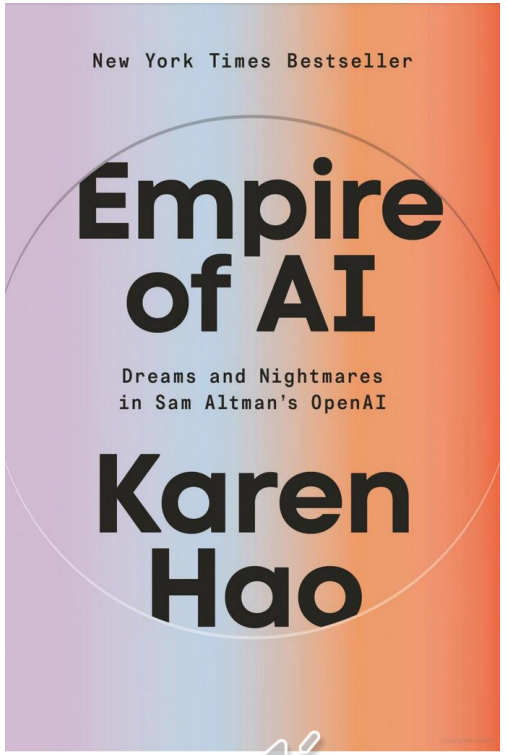
- La majeure partie de cette énergie sera alimentée par les combustibles fossiles. On signale déjà des prolongations de durée de vie des centrales à charbon. Elon Musk a construit son gigantesque supercalculateur Colossus à Memphis, dans le Tennessee, et l'alimente grâce à environ 35 centrales au méthane non autorisées qui rejettent des milliers de tonnes de polluants atmosphériques dans ces communautés.
- Il s’agit donc d’une crise climatique, d’une crise de santé publique et également d’une crise d’eau potable, car bon nombre de ces centres de données s’installent dans des communautés et doivent être refroidis avec de l’eau potable, pas avec autre chose car cela pourrait entraîner la corrosion des équipements et favoriser la croissance bactérienne.
- Et le plus souvent, l’alimentation est eau des datacenters est pirse dans le réseau public d'eau potable, car c'est l'infrastructure pré-existante.
- Google a décidé d'installer un centre de données à Montevideo, la capitale de l'Uruguay. La région était confrontée à une sécheresse historique, à tel point que le gouvernement de Montevideo a commencé à mélanger de l'eau toxique sortant du datacenter au réseau public d'eau potable, simplement pour que de l'eau sorte du robinet des gens. Les personnes trop pauvres pour acheter de l'eau en bouteille étaient contraintes de boire cette eau toxique, et des femmes ont fait des fausses couches.
- Nous assistons donc à l’amplification de nombreuses crises croisées et à la perpétuation de ce paradigme du plus grand modèle, plus de calculs, de données, à tout prix.



Karen Hao, Penguin Press Eds., 2025
https://www.penguinrandomhouse.com/books/743569/empire-of-ai-by-karen-hao/

Q6. Quels déplacements de pouvoir économique et politique pourraient résulter de l'adoption de cette technologie ?

- Dans l’analyse de Karen Hao ou d’autres comme Kate Crawford, cet acaparement des ressources est mené avec la forme d’un empire :
  - ils revendiquent des ressources qui ne leur appartiennent pas et créent des règles suggérant qu’elles leur appartenaient.
  - Ils prétendent avoir une mission civilisatrice et que leur pillage du monde est le prix à payer pour amener tout le monde vers la modernité.



Karen Hao, Penguin Press Eds., 2025
https://www.penguinrandomhouse.com/books/743569/empire-of-ai-by-karen-hao/

En conclusion

- Certaines questions fondamentales sont absentes de la plupart des discussions sur l'IA : quel est le problème auquel ces technologies apportent une solution ? Selon qui ? Comment ce problème pourrait-il être formulé autrement, avec quelles implications pour l'orientation des ressources nécessaires pour y remédier ?
- Quels sont les coûts d'une approche ML basée données, qui les supporte et quelles opportunités perdues en résultent ?
- Comment l'intensification algorithmique pourrait-elle être considérée non pas comme une solution, mais comme un facteur contribuant aux problèmes planétaires croissants – crise climatique, insécurité alimentaire, migrations forcées, conflits et guerres, et inégalités ?
- Comment ces préoccupations sont-elles marginalisées lorsque l'IA, perçue comme une menace existentielle ou une solution miracle, accapare nos ressources et notre attention ?

→ Ces questions restent en suspens tant qu'on considère comme inévitable le dévelppement de l'IA, et qu'on ne prend pas le temps de réflechir aux questions de Postman.

1. **Q1**. Quel problème cette technologie résout-elle ?
2. **Q2**. De qui est-ce le problème ?
3. **Q3**. Quels nouveaux problèmes la résolution de ce problème par cette technologie engendrera t-elle ?
4. **Q4**. Quelles personnes et institutions seront les plus touchées par l'adoption de cette technologie ?
5. **Q5**. De quels changements de langage cette technologie est-elle l'excuse ?
6. **Q6**. Quels déplacements de pouvoir économique et politique pourraient résulter de l'adoption de cette technologie ?
7. **Q7**. Quelles utilisations alternatives (avec possibles effets pervers) pourraient être faites de cette technologie ?

[1] L. Suchman, "The uncontroversial 'thingness' of AI," *Big Data & Society*, vol. 10, no. 2, July 2023.

Quels ont été les éléments de réponse présentés ?

- Le but affiché est bien de produire un système conscient, ou en tout cas aussi proche d'un humain que possible, mais ce n'est pas du tout ce que la technologie (et la science) actuelle permet.
- Identifier et reproduire des correspondances de motifs à partir d'exemples marche pour certaines tâches (ex : reconnaissance de formes), mais ne fonctionne pas pour de nombreuses autres (ex : raisonnement). Cela ne signifie pas que la technologie actuelle est mauvaise ou inutile, mais qu'elle est limitée par les choix simplificateurs sur lesquels elle a été conçue.
- Pourtant le discours dominant reçu par le public, largement créé et orchestré par les producteurs de SIA au premier rang desquels les OGAMAM, prône des usages de cette technologie inadaptés en raison de ses limites intrinsèques.
- Ce discours emploie à mauvais escient des termes relatifs au domaine de la pensée humaine, et crée délibérément avec ces glissements langagiers anthropomorphisant les systèmes et déshumanisant l'esprit humain, un pouvoir suggestif du terme IA servant les intérêts des OGAMAM au détriment de l'intérêt de plusieurs groupes (comme les étudiant·es, les groupes historiquement défavorisés, ou même les entreprises compromettant la qualité de leur production et leur sécurité).
- Les hypothèses sur lesquelles repose le deep learning (identification et exploitation de motifs complexes de corrélation dans les données) sont alignées avec les atouts et l'intérêt des OGAMAM de maintenir leur monopole, dont le plus puissant instrument est le besoin en données et puissance calcul des choix technologiques faits.
- Ces modes de production ont 2 conséquences des plus néfastes : sur les humains victimes de nouvelles formes d'exploitation et face à de nouvelles formes questionables de travail, et sur l'environnement (donc aussi sur les humains).